IMAGINE: Signal and Image Processing Using Streams

Brucek Khailany

William J. Dally, Scott Rixner, Ujval J. Kapasi, Peter Mattson,

Jinyung Namkoong, John D. Owens, Brian Towles

Concurrent VLSI Architecture Group
Computer Systems Laboratory
Stanford University

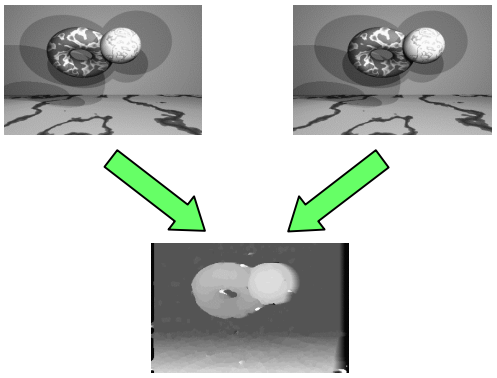
http://cva.stanford.edu/imagine

Imagine: A Programmable Signal and Image Processor

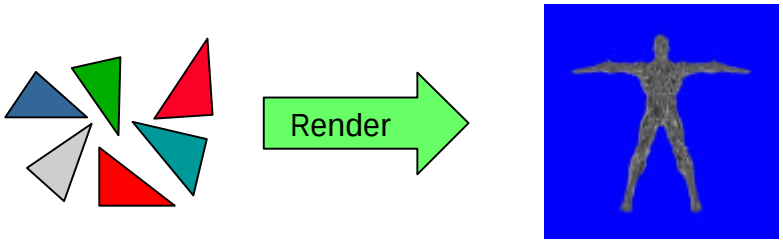
- Motivation
  - Applications poorly matched to conventional architectures
- Key stream architecture features
  - High computational bandwidth (Imagine: 48 on-chip ALUs)
  - Stream register organization
  - Data bandwidth hierarchy
- Performance density of a special purpose processor
  - 0.59 cm^2 CMOS chip, 0.13 μm standard cell, 500 MHz
  - 20 GFLOPS peak performance (40 GOPS fixed point)
  - 10 GFLOPS sustained on several apps
  - > 2 GFLOPS/W, > 5 GOPS/W

Representative Applications

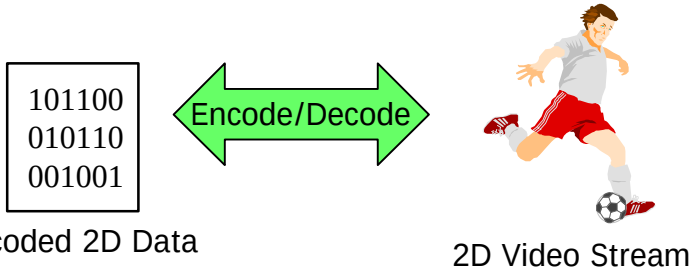
- Stereo Depth Extraction



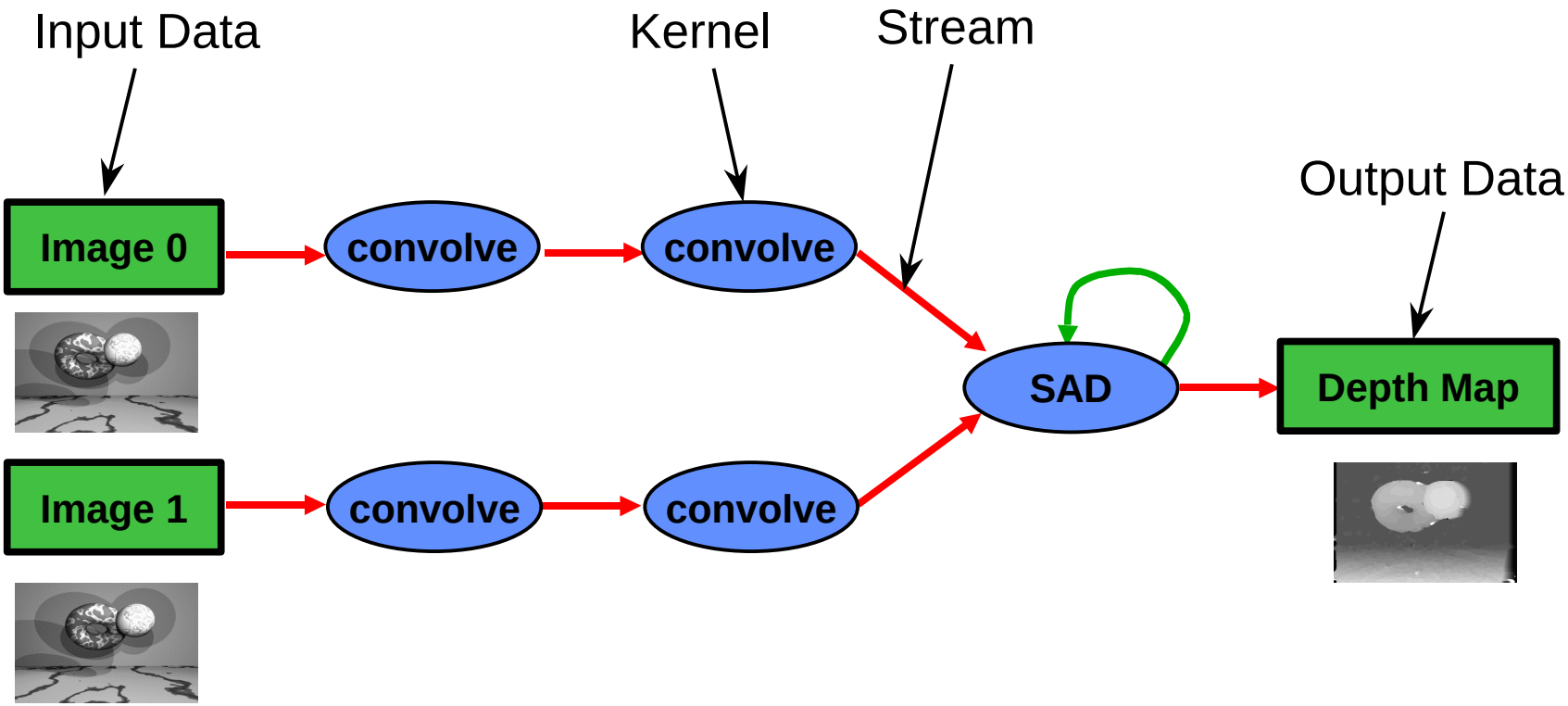
- Polygon Rendering



- MPEG Encoding/Decoding



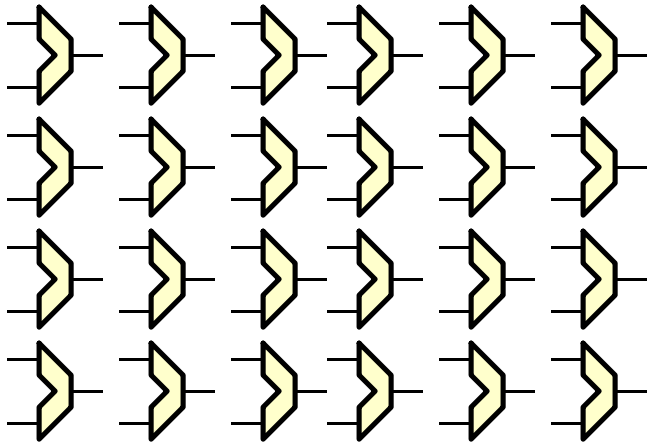
Stream Processing



- Little data reuse (pixels never revisited)
- Highly data parallel (output pixels not dependent on other output pixels)
- Compute intensive (60 arithmetic operations per memory reference)

Characteristics of Media Applications

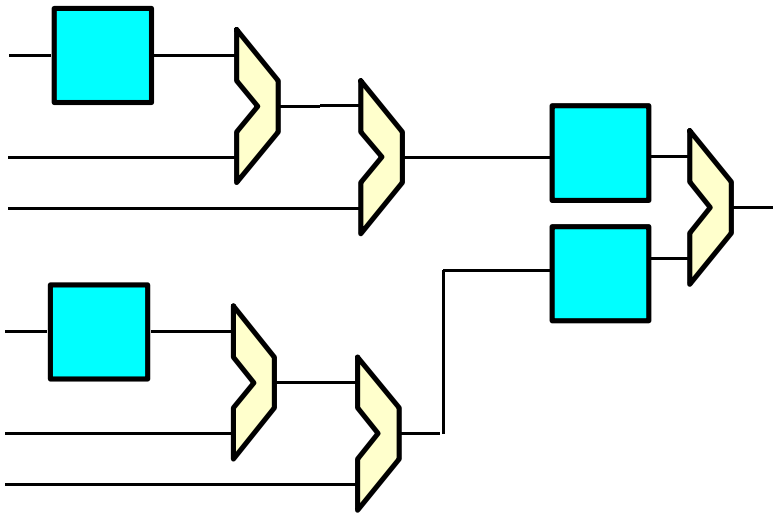
- Poorly matched to conventional architectures
  - Instruction-Level Parallelism
  - Caches
  - Few arithmetic units
- Well-matched to modern VLSI technology
  - Lots (100's - 1000's) of ALUs fit on a single chip



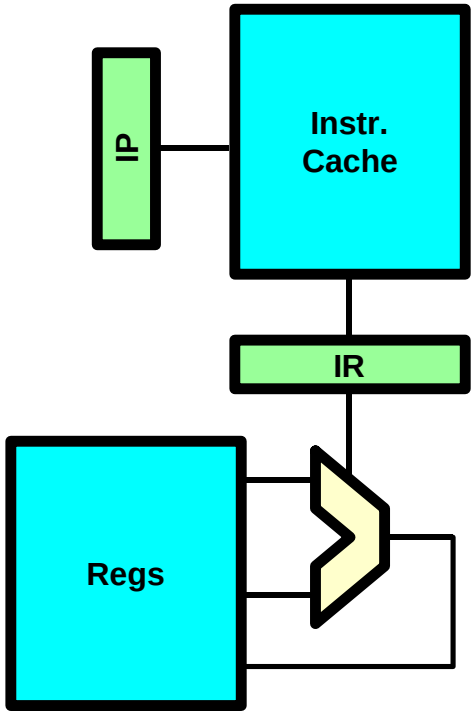
  - Communication bandwidth is the scarce resource

Communication Bandwidth: Care and Feeding of ALUs

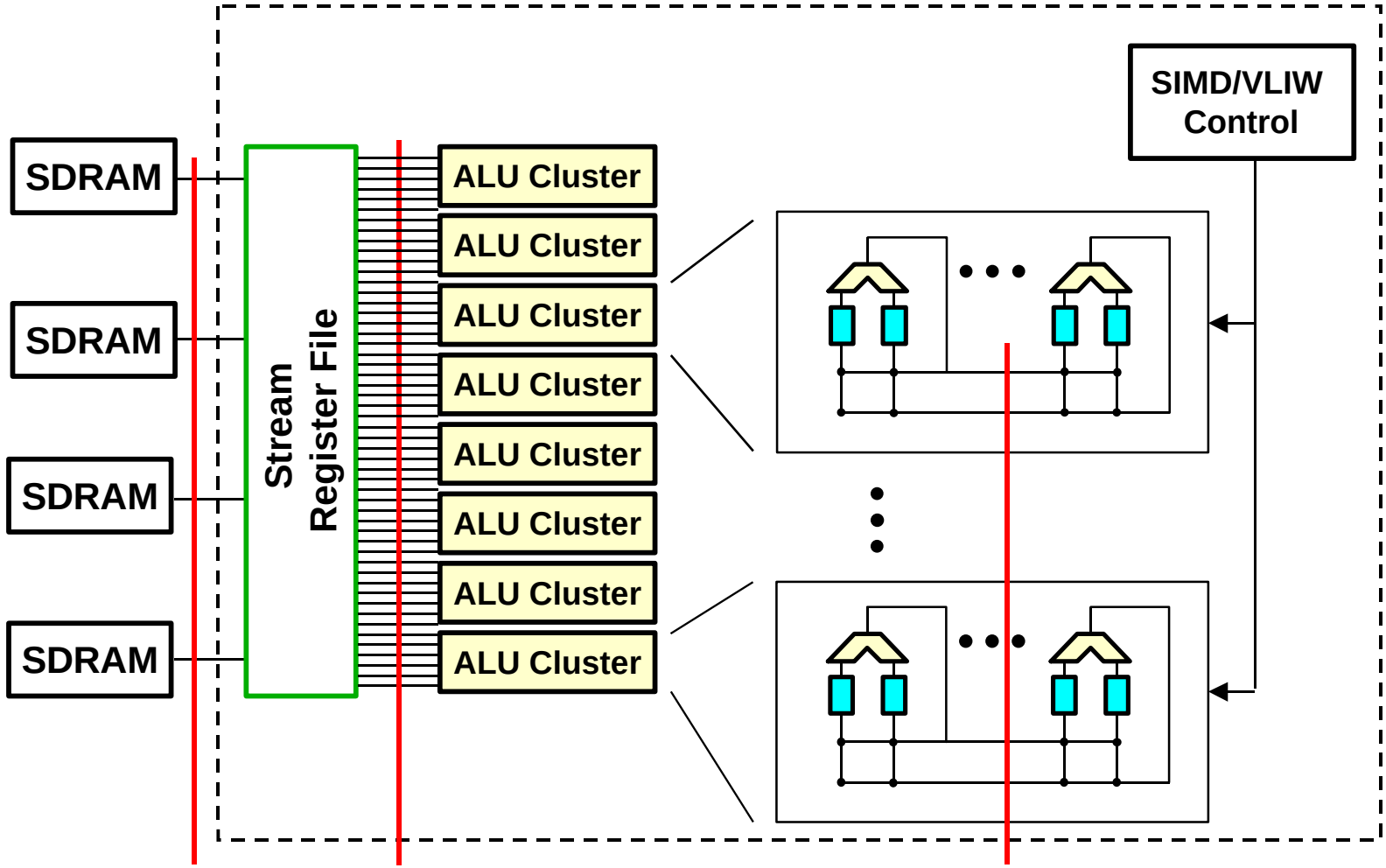
Special-Purpose Processors:
ALUs fed by dedicated wires/memories



General-Purpose Processors:
'Feeding' Structure Dwarfs ALUs

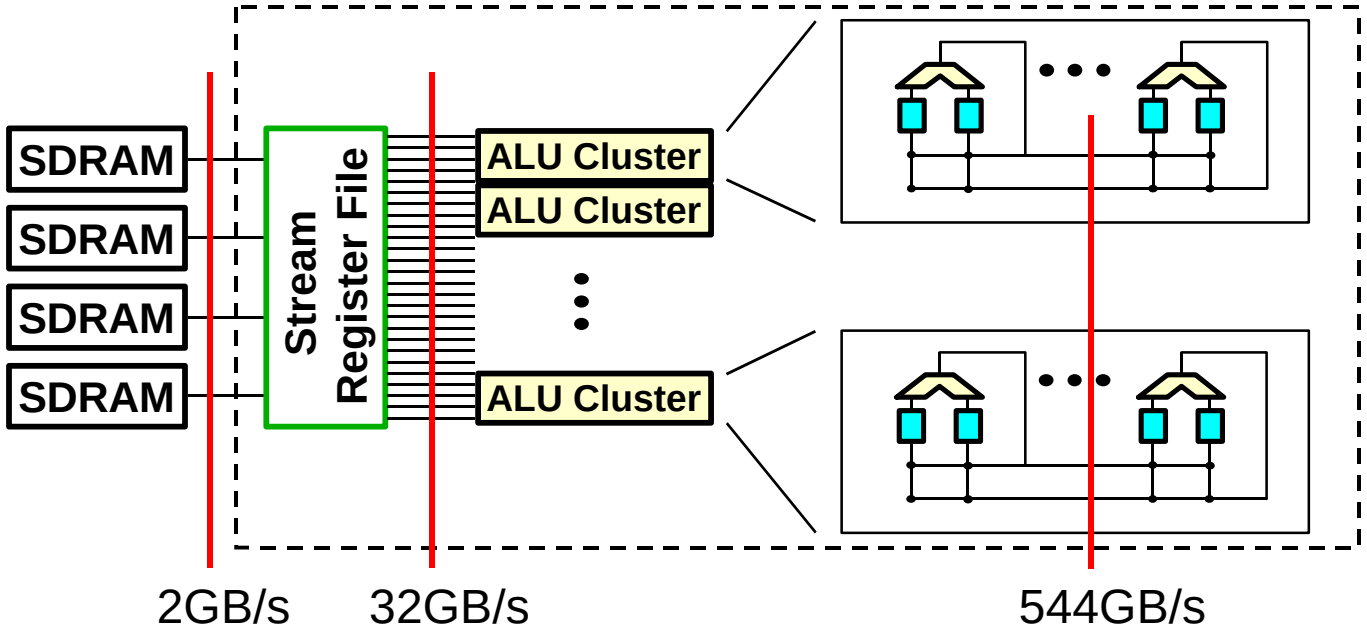


Stream Architecture Provides Data Bandwidth Hierarchy



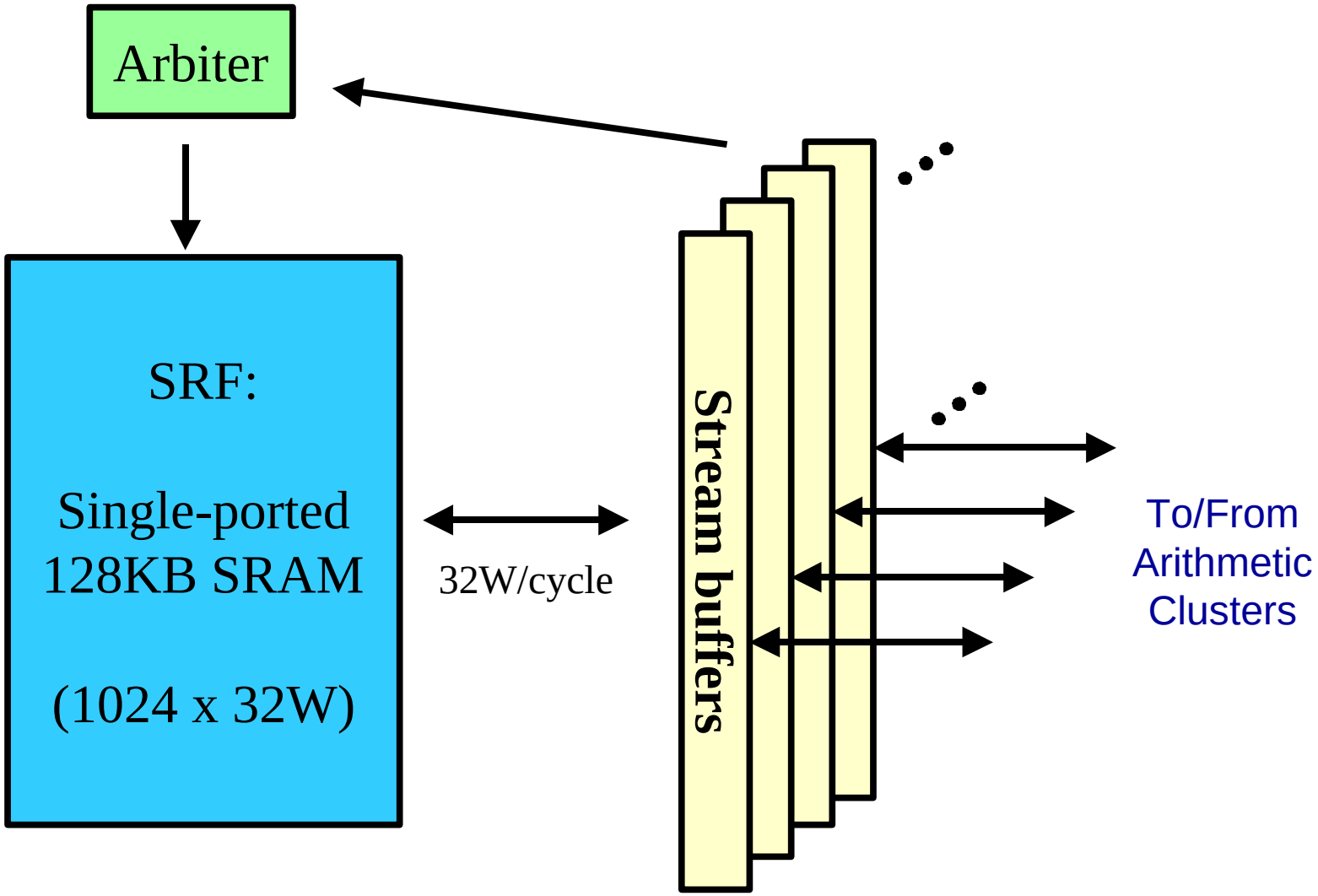
Peak BW: 2GB/s 32GB/s 544GB/s

Application Data Bandwidth Usage

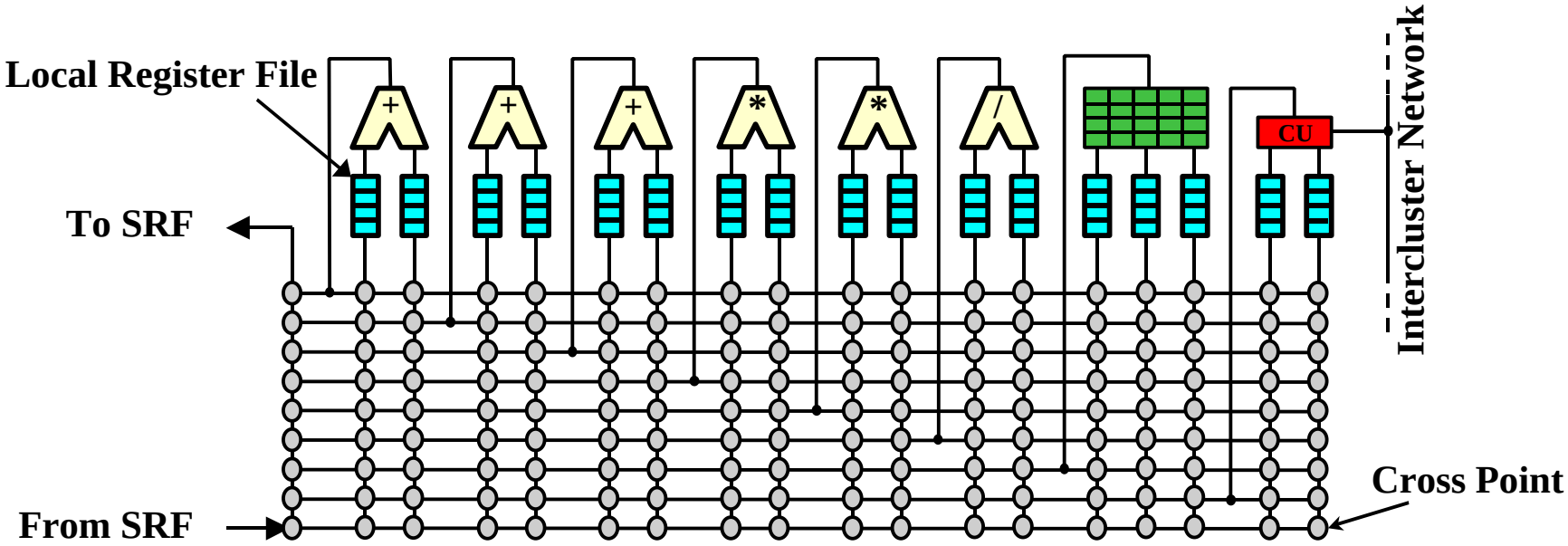


	Memory BW	*Global RF BW*	*Local RF BW*
Depth Extractor	0.80 GB/s	18.45 GB/s	210.85 GB/s
MPEG Encoder	0.47 GB/s	2.46 GB/s	121.05 GB/s
Polygon Rendering	0.78 GB/s	4.06 GB/s	102.46 GB/s
QR Decomposition	0.46 GB/s	3.67 GB/s	234.57 GB/s

Stream Register File: Details

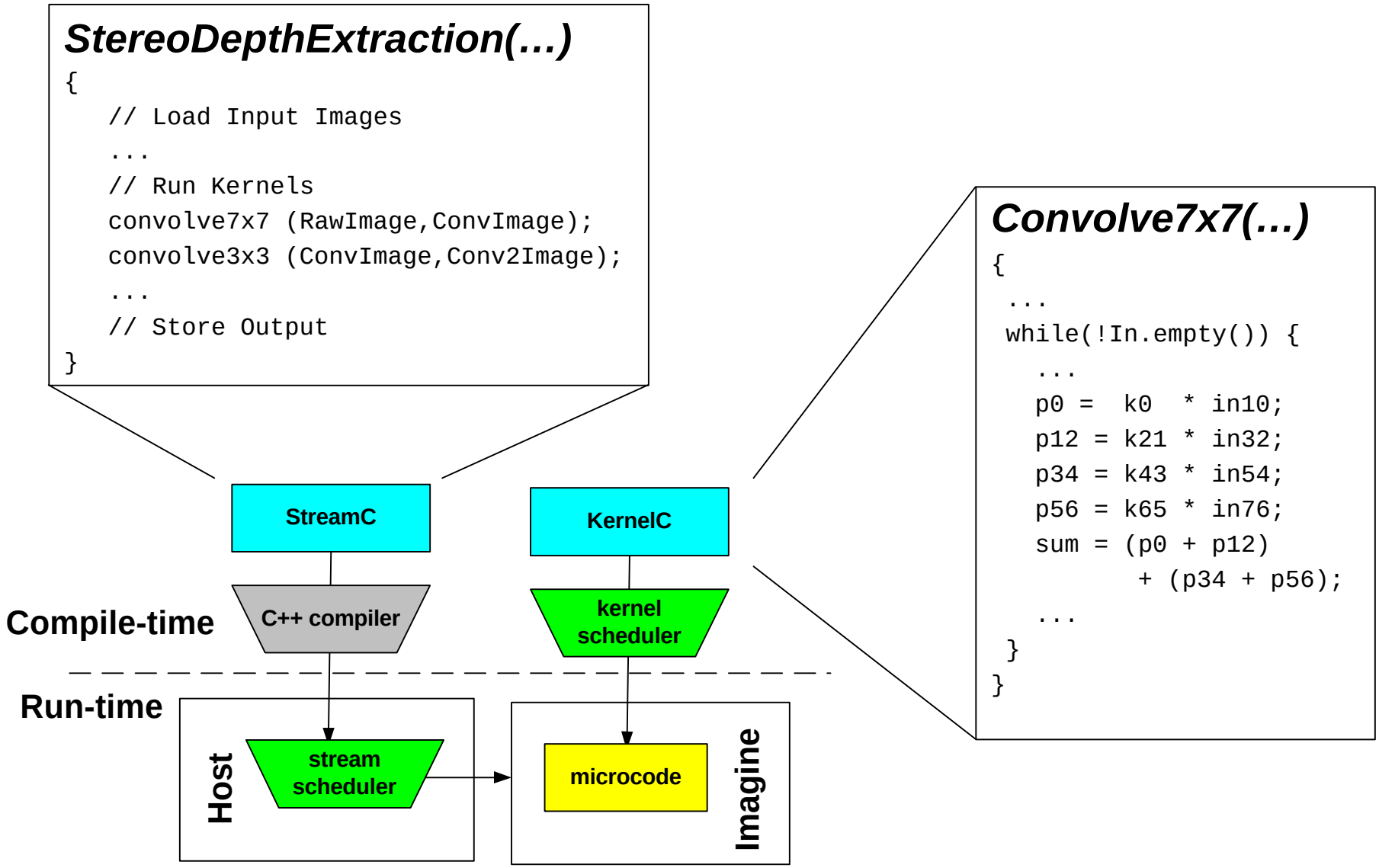


Arithmetic Cluster: Details

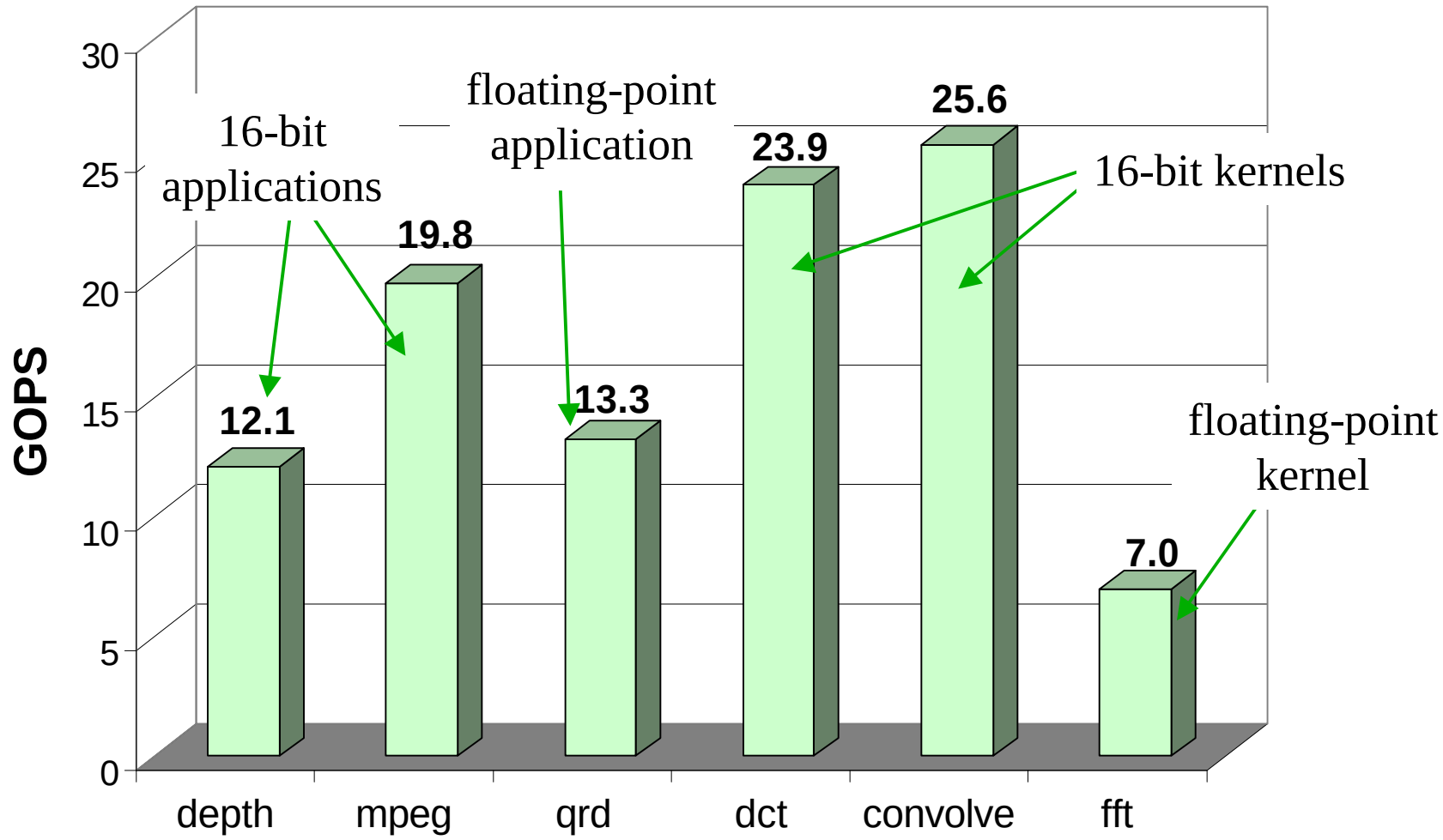


- Units support floating-point / 32-bit / dual 16-bit / quad 8-bit instructions
  - 4-cycle pipelined FMUL,FADD,FSUB,FTOI,ITOF,FFRAC
  - 17-cycle FDIV (pipelined for 1 FDIV every 7 cycles)

Imagine Programming Environment

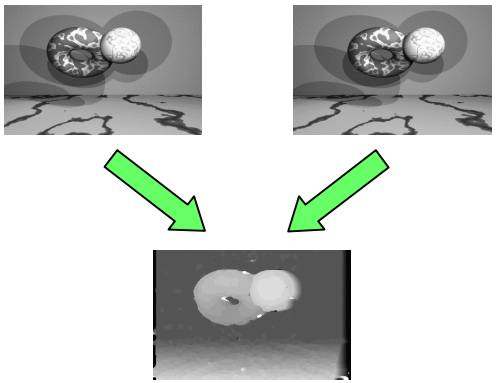


Performance

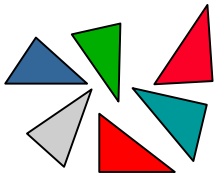


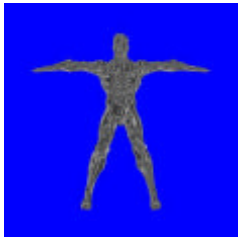
Sustained Application Performance

- Stereo Depth Extraction
  - 320x240 8-bit grayscale
  - 200 frames/second



- Polygon Rendering
  - 4.5 Million Vertices/sec
  - 5.1 Million Pixels/sec

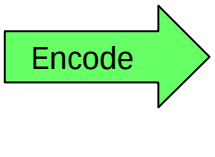


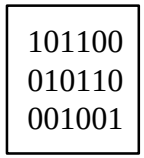


SPECviewperf ADVS benchmark (unlit)

- MPEG Encoding
  - 720x486 24-bit color
  - 120 frames/second

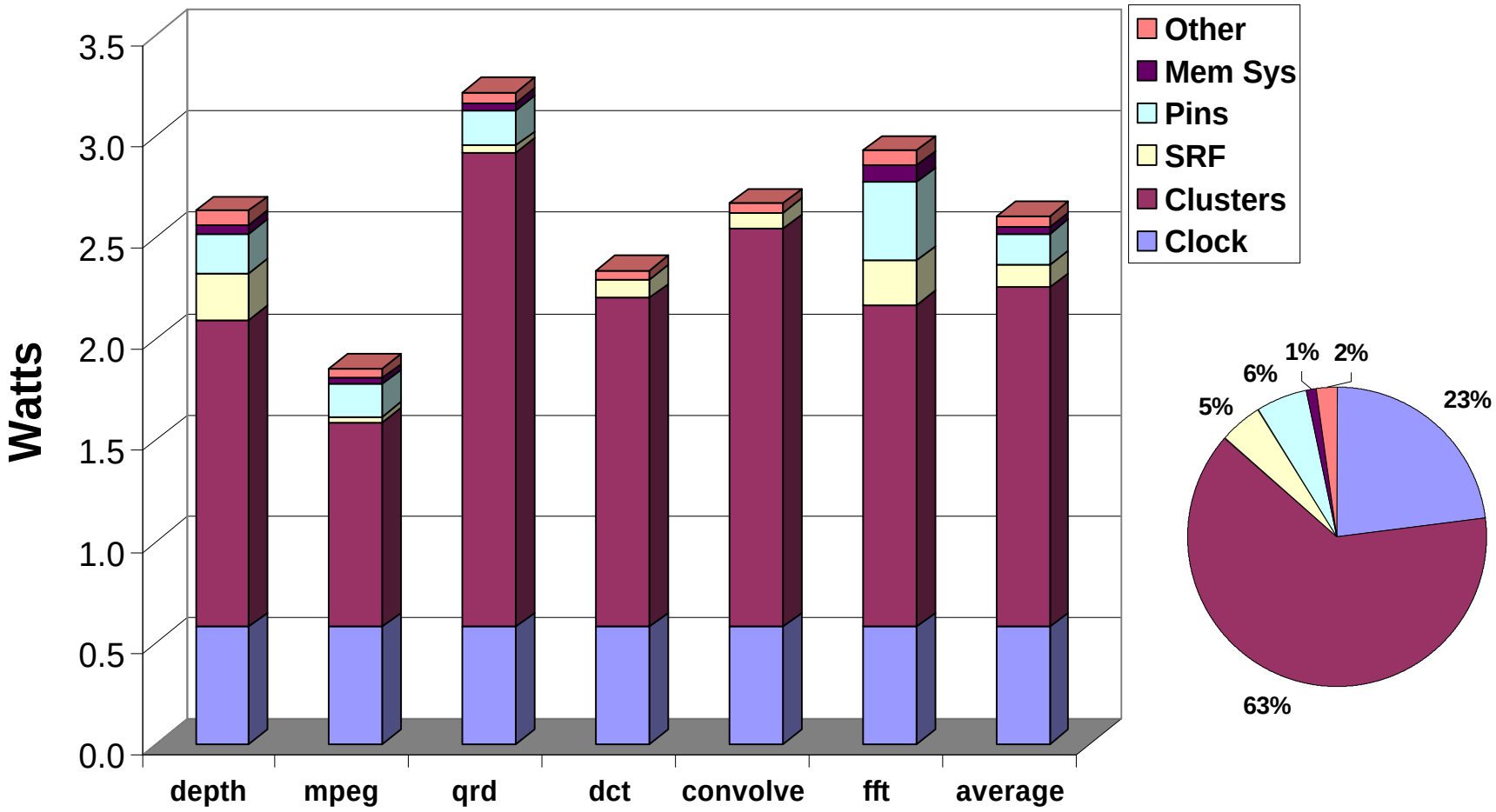
2D Video Stream





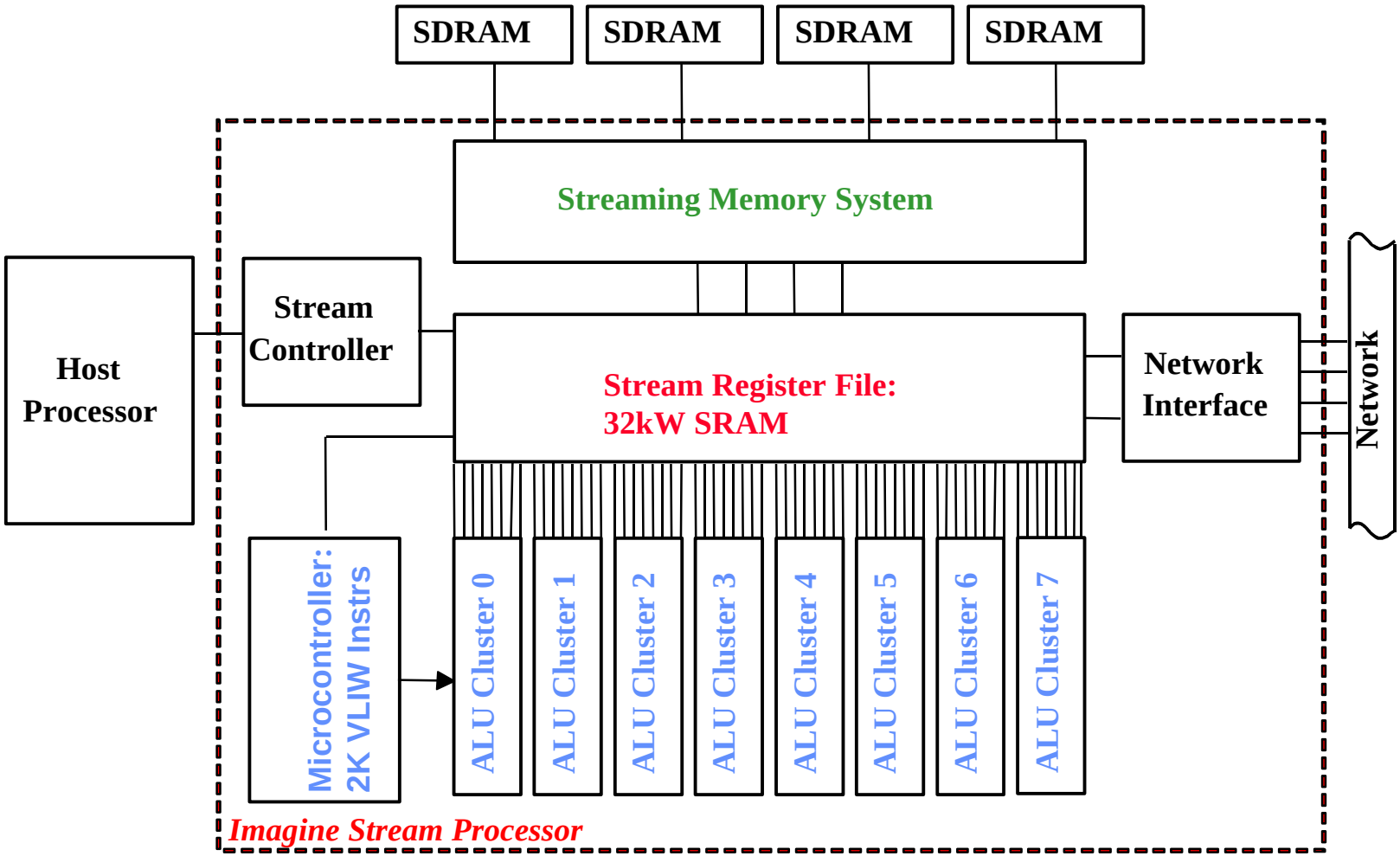
Encoded 2D Data

Power Estimates



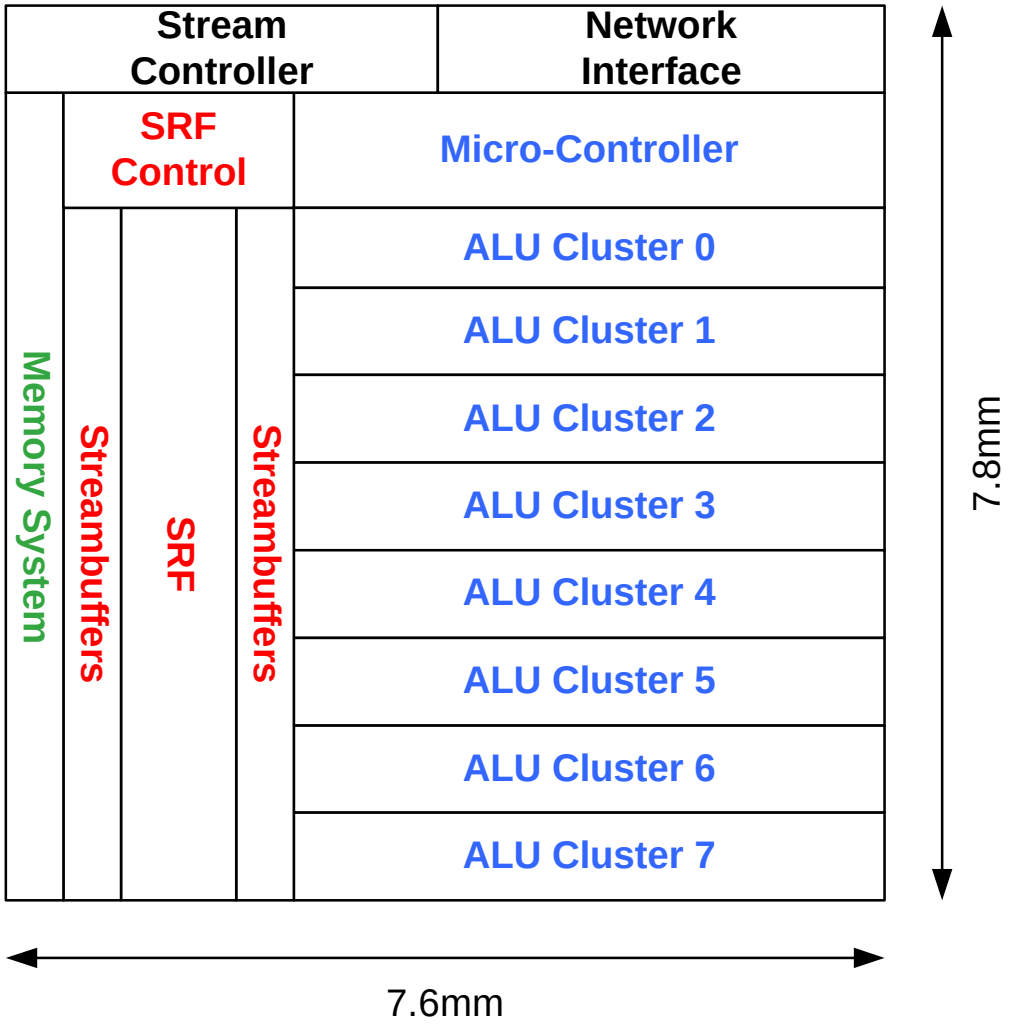
GOPS/W:	4.6	10.7	4.1	10.2	9.6	2.4	6.9

The Imagine Stream Processor



Imagine Floorplan

- 22 million transistors
- 500 MHz
- TI GS30KA:
  - 0.15 µm L_{drawn}
  - 0.13 µm L_{eff}
  - CMOS process



VLSI Implementation: 22M Transistors with 7 grad students

- Stream architecture reduces VLSI design complexity
  - Modularity / Replication
  - Long wire delays converted to explicit communications
    - Exposed to microarchitecture, software
- Design methodology
  - Standard ASIC flow with forced placement of datapaths
    - Bitslice Verilog
    - Improved area, delay
  - Pre-placement wire length estimates
    - Reduce design iterations

Status

- Imagine team accomplishments
  - Cycle-accurate simulator
  - Software tools
  - Completed synthesizable Verilog
  - Arithmetic units implemented in standard cells
- Industrial partners
  - Texas Instruments: Fab
  - Intel
- Future work
  - Circuits/Logic: expected completion 9/15/00
  - Tapeout: expected Q4/2000

Summary

- Key stream architecture features
  - Stream register organization
  - Data bandwidth hierarchy
- Performance density of a special purpose processor
  - 10 GFLOPS sustained on several apps
  - >2 GFLOPS/W, >5 GOPS/W
- VLSI Implementation
  - Validate architectural concepts
  - Develop experimental prototype